Escape your echo chamber! The World News Organizer

Project Description:

A. Overview

This is an application that shows international news and headlines from reliable news sources such as BBC, CNN, and Al Jazeera. Other than the news itself, it will also show topics related to the news for the users to broaden their understandings about the issue reported.

B. Motivation

People nowadays have gradually been considered indifference towards global issues and world topics. There is a term that has become increasingly popular in recent years called "同溫層", resembling the term "echo chamber" in English, which refers to an environment in which a person encounters only beliefs or opinions that coincide with their own. The best way for a person to leave his or her echo chamber is to read news about the world. However, local newspapers and media in Taiwan usually cover only little information about the world and are considered biased. Therefore, I came up with the idea to gather international headlines from trusted and reliable sources and display them all at a time, the user could then get a grasp of world issues at a glance. Other than this, adding related topics can also help users to understand what are related to the news.

Project Planning:

The entire project can be broken into three parts: **A.Data collection**, **B.Data analysis**, and **C.Data Representation**.

A. Data collection (Using NewsAPI and Web Crawling):

The data that is going to be collected are world news headlines and its respective contents (photos and links). The news sources are intended to be from big corporations such as BBC, Reuters, CNN, The Japan Times and etc. Many of those can be accessed through the **NewsAPI**, which would return data in the form of **json** that can be easily accessed in python. However, the NewsAPI does not include information from all news sources (e.g. The Japan Times), and there is also no API to get data, therefore we would have to use **BeautifulSoup** to crawl those websites in order to get data.

B. Data analysis (Using Text Analytics API and PyTrends):

To show topics that are related to the respective news/headline, we would first have to analyze the data and extract keywords and information from it. To do so, I've decided to make use of the **Text Analytics API** from Microsoft Azure. This API is capable of identifying key phrases and entities of the provided text.

Then, we would use **PyTrends**, an unofficial API for Google Trends, which can show information such as "related topics" or "interest over time" of the given keyword. We can plug in our key phrases into PyTrends so as to achieve the list of related topics.

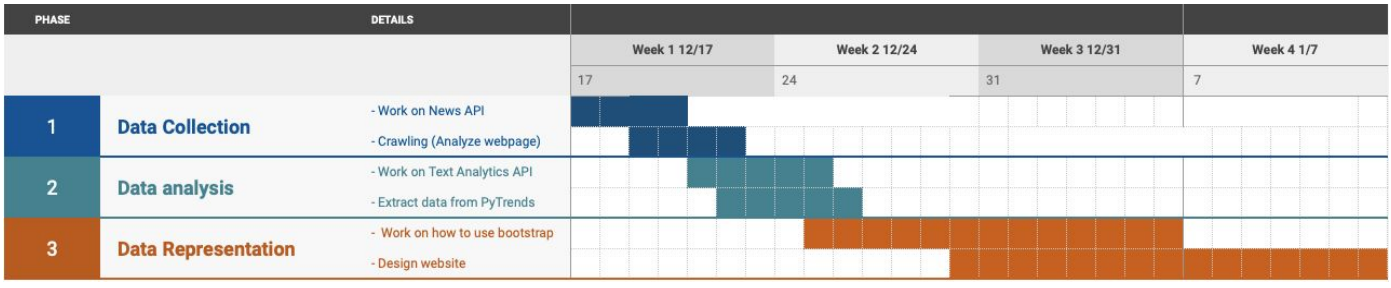
C. Data Representation (Simple HTML page):

After obtaining, analyzing and arranging the data, the last part is to represent the data for the user to see. I am planning to present the information using an HTML page. The page is going to show a few blocks, each showing the title, image, and description of a single piece of news. This can be done by using Django or Bootstrap.

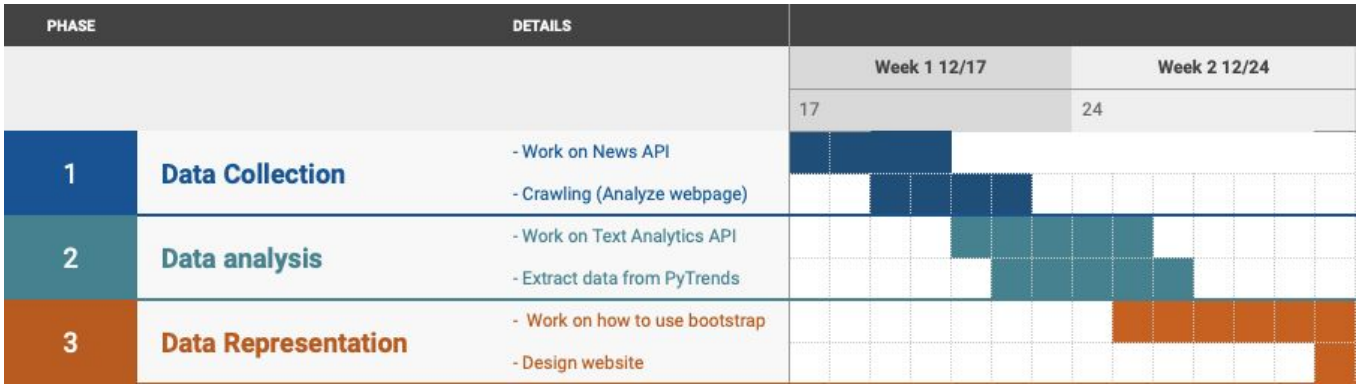
User Interaction:

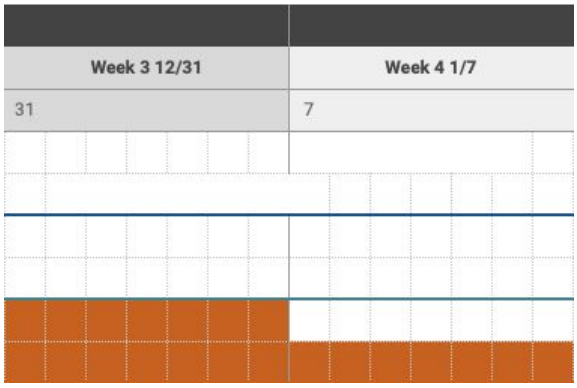
When opening the application, it is going to give the user a list of 10 news sources to choose from. The user can choose 3 of them by entering the respective code of the source. After confirming the sources, the application then generates a webpage that shows the information. The user can scroll through the page to read the news, and the title of each article can be clickable, which will lead to the original website of the news.

Timeline:



Larger version:





Progress Update 1:

A. What I've Done

a. Successfully grab headlines from each source with NewsAPI

The news sources are each respectively:

i. Reuter's
ii. BBC
iii. The Wall Street Journal
iv. The Associated Press
v. Al Jazeera
vi. The New York Times
vii. TIME
viii. CNN

b. Crawl each website for keywords

```
<meta name="description" content="After evangelical publication
Christianity Today published a blistering editorial on what it
called U.S. President Donald Trump's "grossly immoral characte">
<meta name="robots" content="noodp">
<meta name="keywords" content="U.S., abortion, LGBT, racism,
Donald Trump, evangelicals, billy graham, gay marriage"> = $0
```

For each individual news, there are tags with keywords in the HTML file (like the one above). I used BeautifulSoup to crawl the data.

c. Wrote a simple html to show the obtained information

Thousands Flee to Shore as Australia Fires Turn Skies Blood Red

Source: Isabella Kwai

Description: The countryâ€™s east coast was dotted with apocalyptic scenes on the last day of the warmest decade on record in Australia.

Keywords: Fires-Australia-Victoria-New South Wales-

Read more

Iraq Protesters Break Into U.S. Embassy Compound in Baghdad

Source: Falih Hassan, Ben Hubbard

Description: Demonstrators marched on the compound, breached the outer wall and lit fires. Tensions are high after American airstrikes killed 24 members of an Iraqi militia backed by Iran.

Keywords: Diplomats Embassies and Consulates-Baghdad-Iraq-US-Iran-Civil Unrest-

Read more

Carlos Ghosn, Fallen Nissan Boss, Has Unexpectedly Left Japan

Source: Emily Flitter

The final version would be something like this.

B. Changes on the final project plan

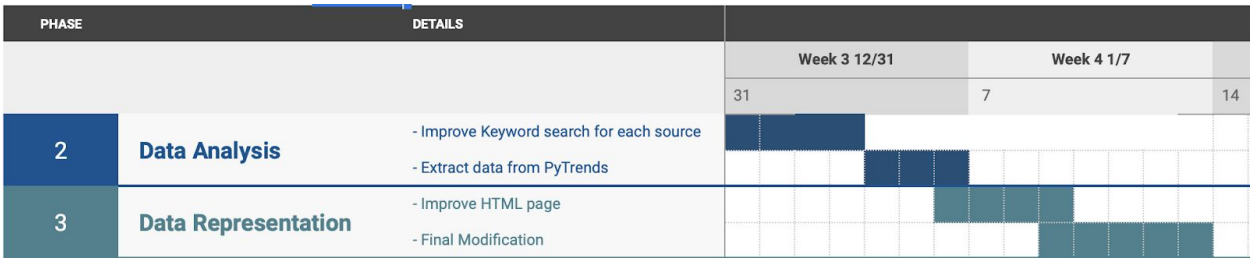
a. Not using Microsoft Azure Text Analytics API

Although the API is really powerful, the keyword results are not in the form that I was expecting. In addition, I found out that I could directly get the keyword data from crawling the news webpage, therefore I chose not to use the Text Analytics API.

b. Change from 10 news sources to 8 news sources

10 news sources are a bit too much.

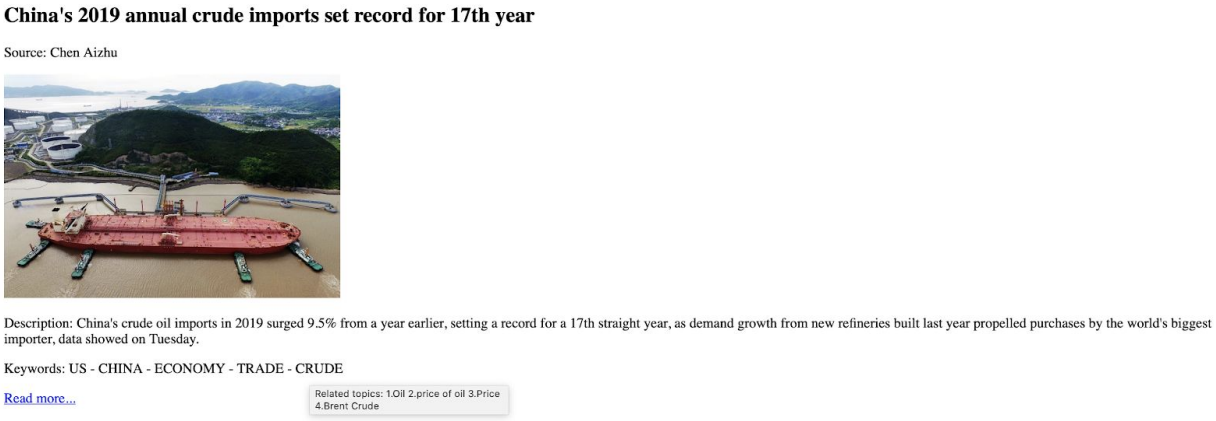
Timeline:



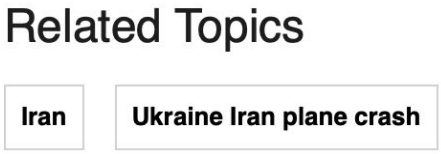
Progress Update 2:

Changelog

A. **Complete the Related topics feature for each individual news**
   By using PyTrends API, we can achieve the related topics of a specific keyword in a *pandas DataFrame* format. The related topics can be accessed by hovering your mouse on top of the keywords, an example is shown below:

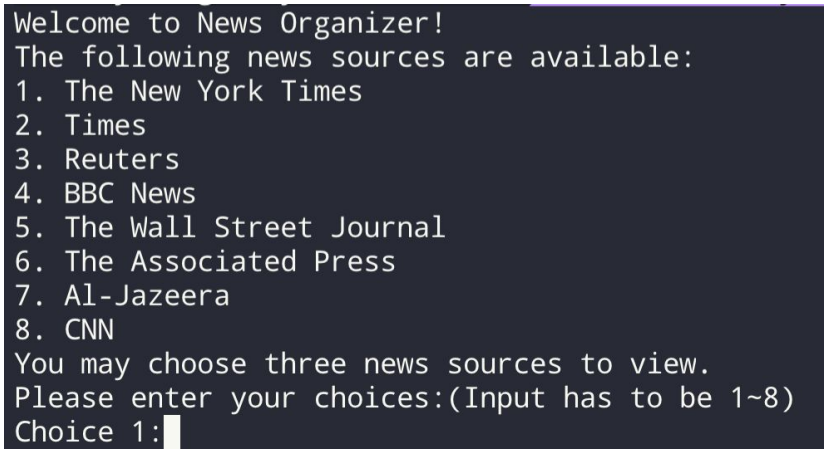


B. **Improve Keywords Crawling**
   At first, the keywords of the websites were achieved by crawling the metadata that is in the header of a webpage. But some websites have their own organized keywords, for example, BBC news has its own related topics section:



   Therefore, I have chosen to individually crawl the keywords for the websites.

C. **Integrate all sources together**



Added a choice interface, the user can choose 3 news sources from the 8 sources provided.

D. **Auto open html file after grabbing data**

By using the `subprocess` library in python, the script will auto open the generated html file with the default browser of the system.

E. **Fixed html display issues with escape sequences**

Characters like apostrophes cannot correctly display if directly written into document, therefore, they are replaced with html escape sequences for it to display correctly in html file.

F. **Divided the file into main.py and function.py**

Divided the entire file into two portions, for easier maintenance.

Known Issues:

A. **News sources from NewsAPI are not stable**

Some sources will be unavailable from time to time due to unknown issues from NewsAPI. For example, BBC disappeared for an afternoon (that is, does not return any data when requested) when I was testing, but then reappeared later. And The New York Times suddenly became unusable on the day before the final demo, I'm not sure whether it would be usable later on. Therefore, The New York Times is unusable for now.

B. **There is an unknown limited of requests with PyTrends**
This problem is stated in the "Caveats" section in the **PyTrends Documentation**, since this is not an official API from Google. I have used a timeout function to prevent the problem, but this does fix the problem absolutely. Therefore, make sure don't run the program too many times in a short amount of time, so that the Pytrends would not be unusable.

C. **It takes to much time to run the program**
The overall time to run the program is about one and a half minute, and this is due to the PyTrends API and how I implemented it, because I have to obtain the related topic for every single keyword of the news.

How to install

- **Requirements:**
  - Python 3.3+
  - Requests, Pandas
  - BeautifulSoup4
  - NewsAPI
  - PyTrends

- **Installation (MacOS)**
The python packages can all be installed using **`pip`** under terminal, the commands for each package are shown below:

  - Requests
    ```
    pip3 install requests
    ```

  - Pandas
    ```
    pip3 install pandas
    ```

  - BeautifulSoup4
    ```
    pip3 install beautifulsoup4
    ```

  - NewsAPI
    ```
    pip3 install newsapi-python
    ```

  - PyTrends
    ```
    pip3 install pytrends
    ```

How to run

Step 1:

There are two python files in this project, **`main.py`** and **`functions.py`** respectively. In order to run the project,

For MacOS:

Under the same path of the files in terminal, type

```
python3 main.py
```

For Windows:

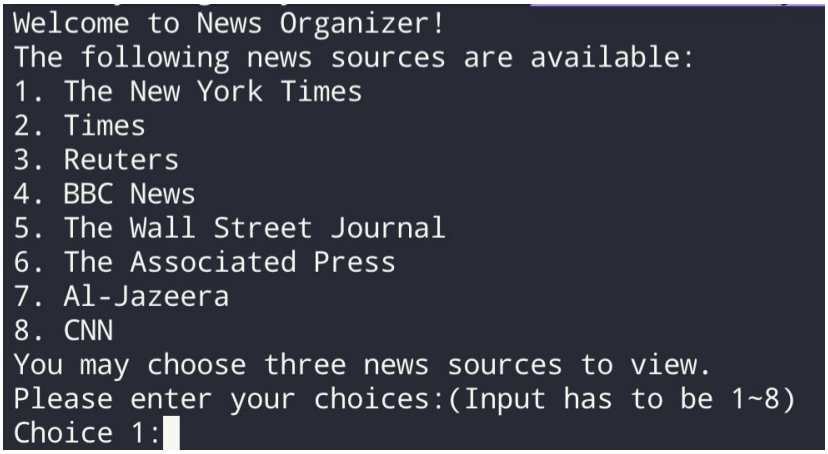
In CMD, type

```
py main.py
```

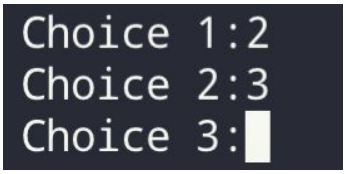
To run the file.

Step 2:

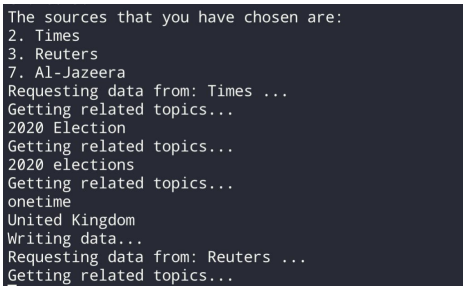
After running the file, the program will prompt you to choose 3 news sources from the 8 sources below as shown below:



Enter your choice and press enter, the program will continue to prompt you.



After entering three choices, the program will start to request data:



Step 3:
After requesting data, the program will generate an HTML file **`pages.html`**, and automatically open it with the default browser of your system:

With No Clear Frontrunner, 6 Candidates Head Into The Democratic Debate in Iowa

Source: Lissandra Villa / Des Moines, Iowa

Description: Ahead of the Democratic debate in Des Moines, Iowa there are four leading candidates: Bernie Sanders, Elizabeth Warren, Pete Buttigieg and Joe Biden.

Keywords: 2020 Election

Read more...

Cory Booker Drops Out of 2020 Presidential Race

Source: Philip Elliott

There will be 3 headlines from each chosen source, 9 articles in total.

Description: At least 124 people killed as severe weather triggers avalanches and landslides in Pakistan, India and Afghanistan.

Keywords: Pakistan - Environment - Afghanistan - Asia

Read more...

Related topics: 1.Price 2.Pakistan national cricket team 3.India 4.Indian people

If you hover your mouse on a keyword, the page will show related topics of the respective keyword.
If you are interested in one article and want to read more, you can click on the Read more… and it will lead to the news website.